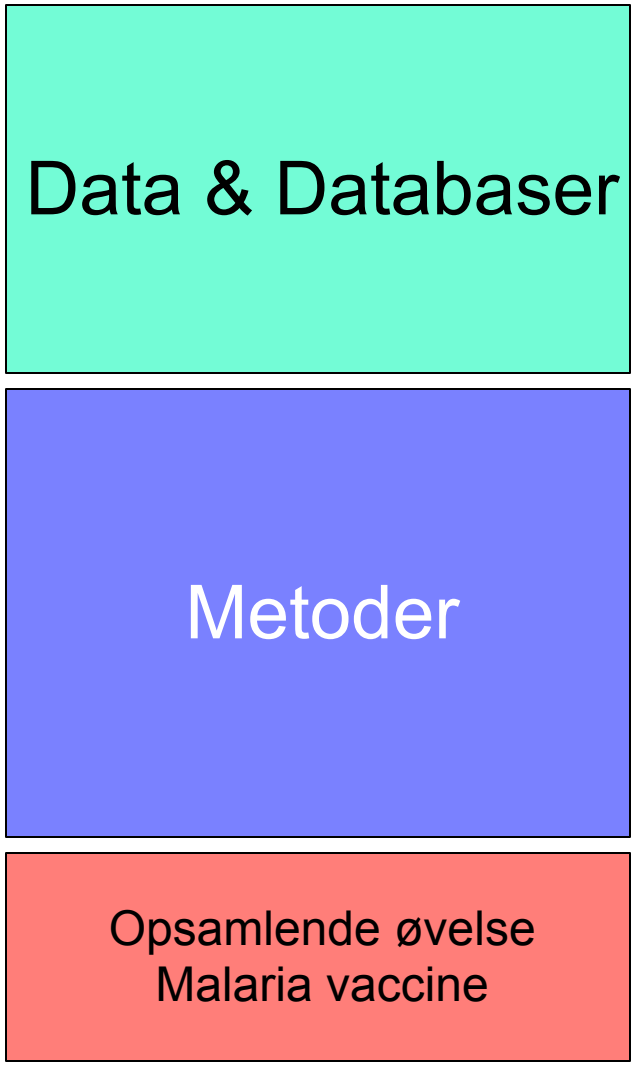
Introduktion til Bioinformatik

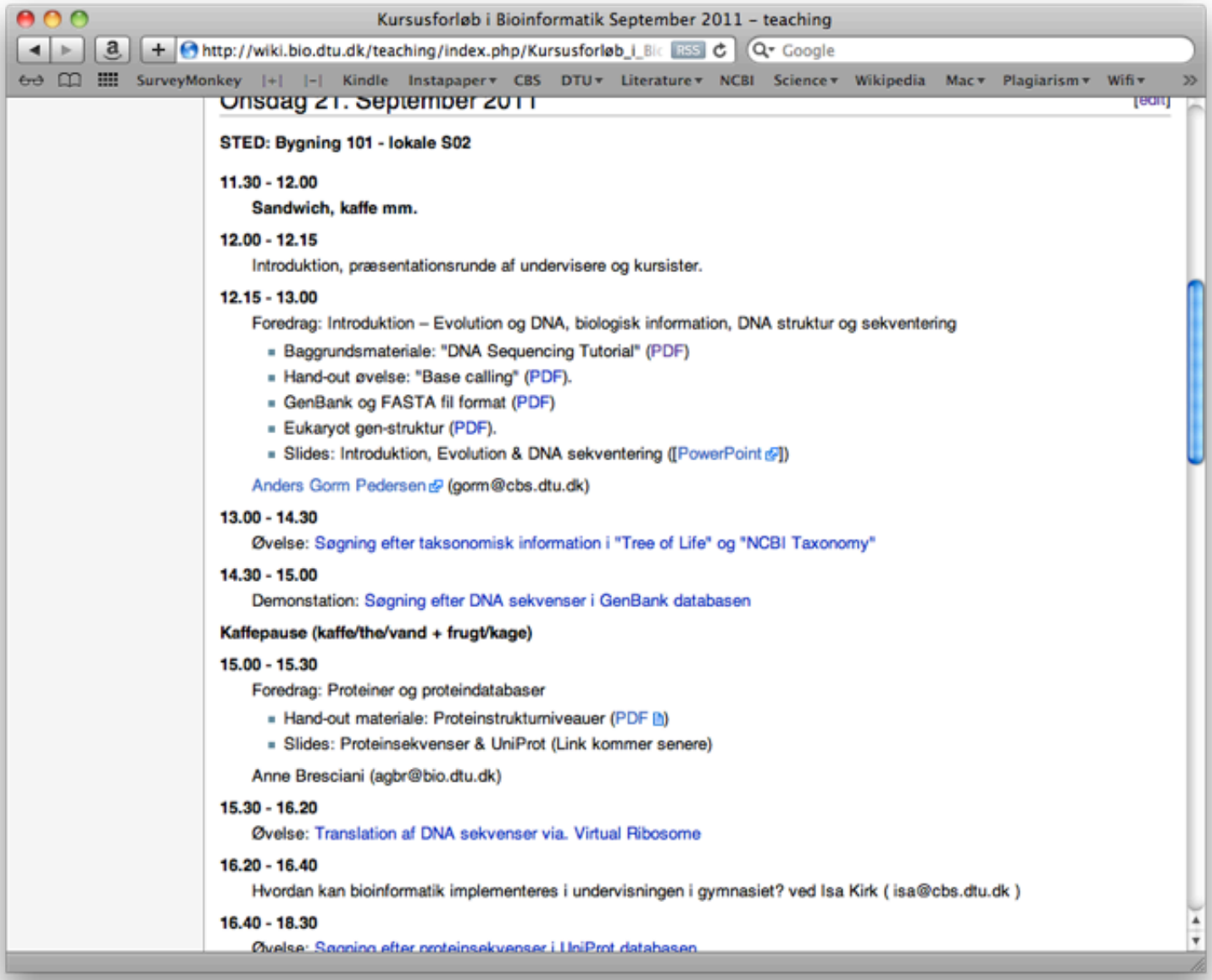
Oversigt



- Taksonomi
- DNA
- Protein
- Protein struktur

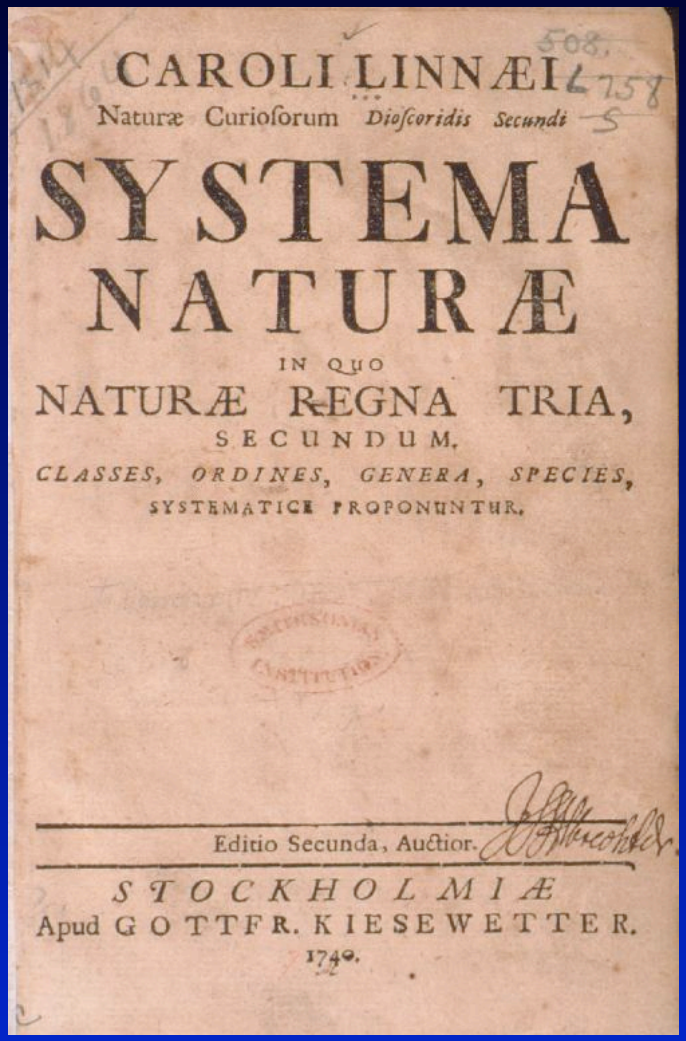
- Alignment
  - Pairwise + Multiple
- BLAST (søgning)
- Fylogenetiske træer
- PyMOL (3D visualisering)

Kursusplan på vores wiki



Classification: Linnaeus

Carl Linnaeus
1707-1778



CAROLI LINNÆI
Naturæ Curioſorum *Dioſcoridis Secundi*

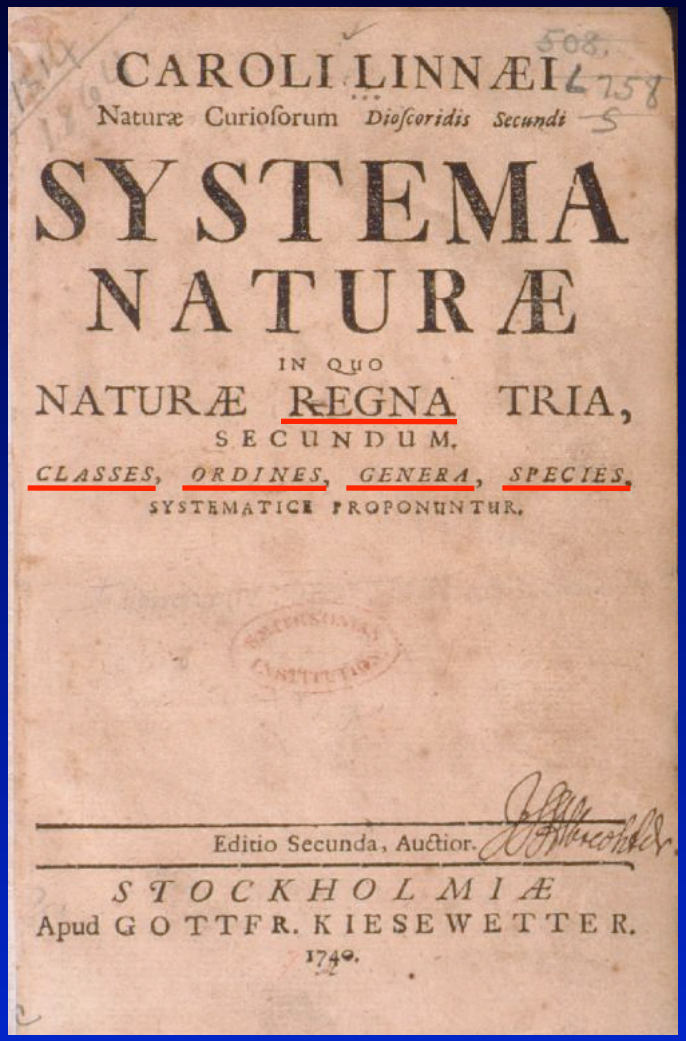
SYSTEMA NATURÆ

IN QUO
NATURÆ REGNA TRIA,
SECUNDUM,
CLASSES, ORDINES, GENERA, SPECIES,
SYSTEMATICE PROPONUNTUR.

Editio Secunda, Auctior.

STOCKHOLMIÆ
Apud GOTTFR. KIESEWETTER.
1740.

Classification: Linnaeus



CAROLI LINNÆI

Naturæ Curioſorum *Dioſcoridis Secundi*

SYSTEMA NATURÆ

IN QUO

NATURÆ REGNA TRIA,

SECUNDUM

CLASSES, ORDINES, GENERA, SPECIES,

SYSTEMATICE PROPONUNTUR.

Editio Secunda, Auctior.

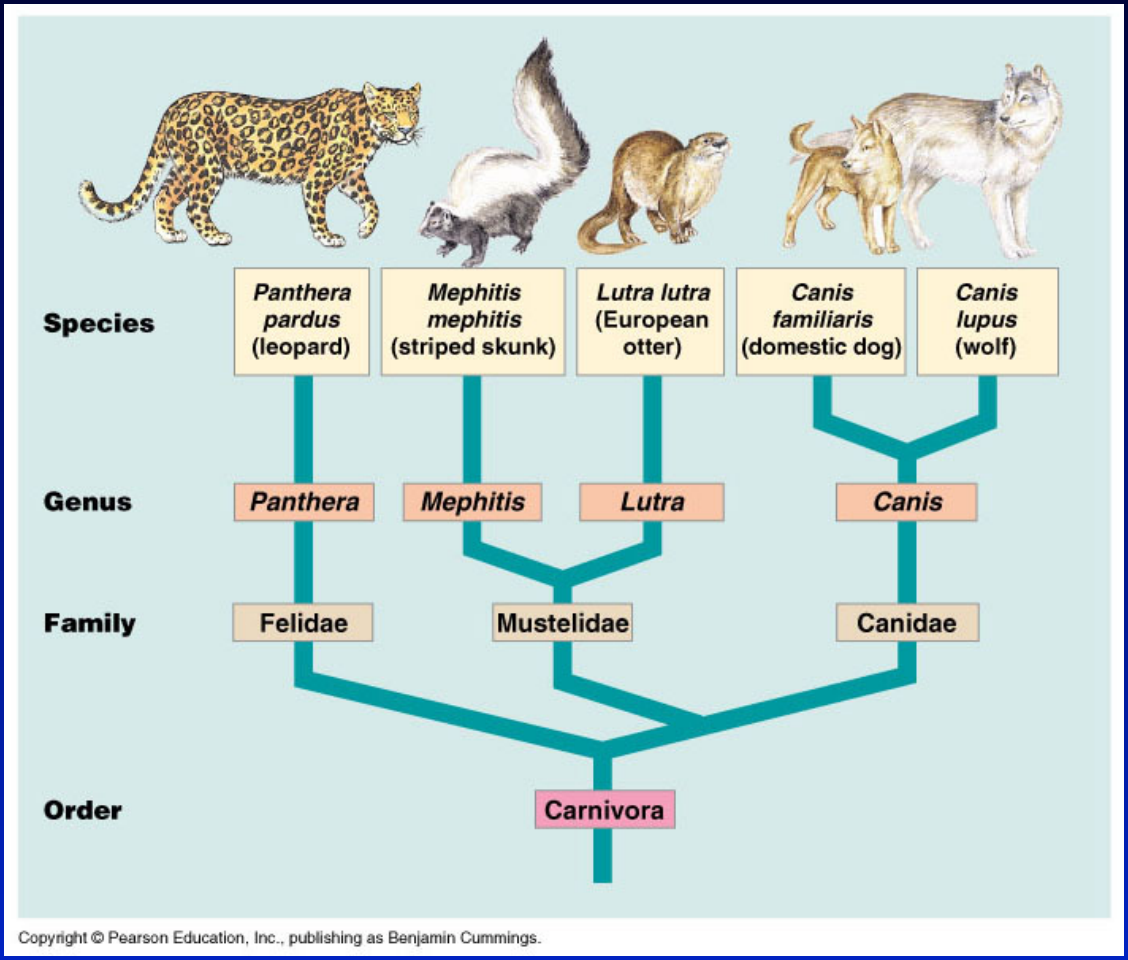
STOCKHOLMIÆ

Apud GOTTFR. KIESEWETTER.

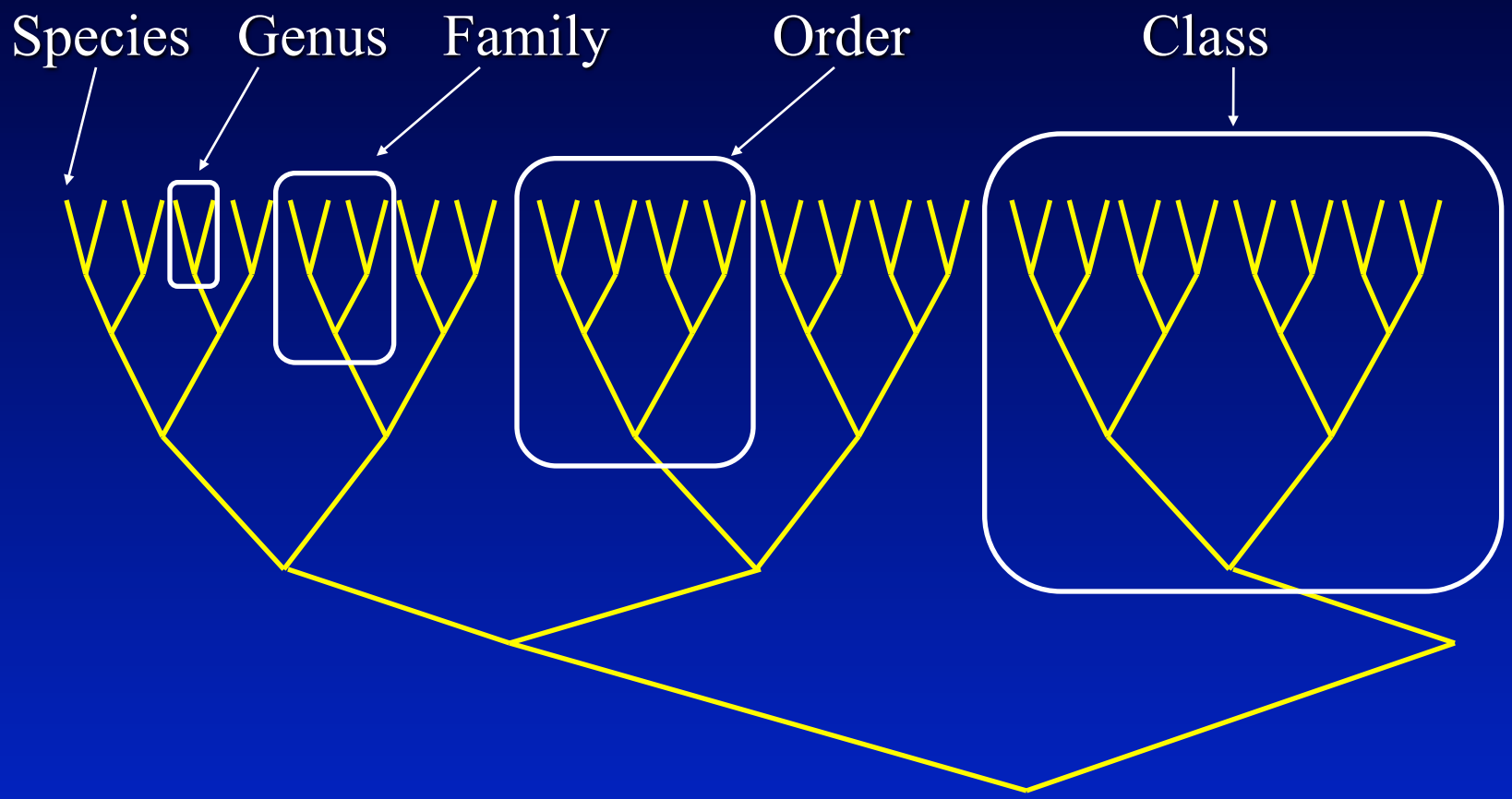
1740.

- Hierarchical system
  - Kingdom (Rige)
  - Phylum (Række)
  - Class (Klasse)
  - Order (Orden)
  - Family (Familie)
  - Genus (Slægt)
  - Species (Art)

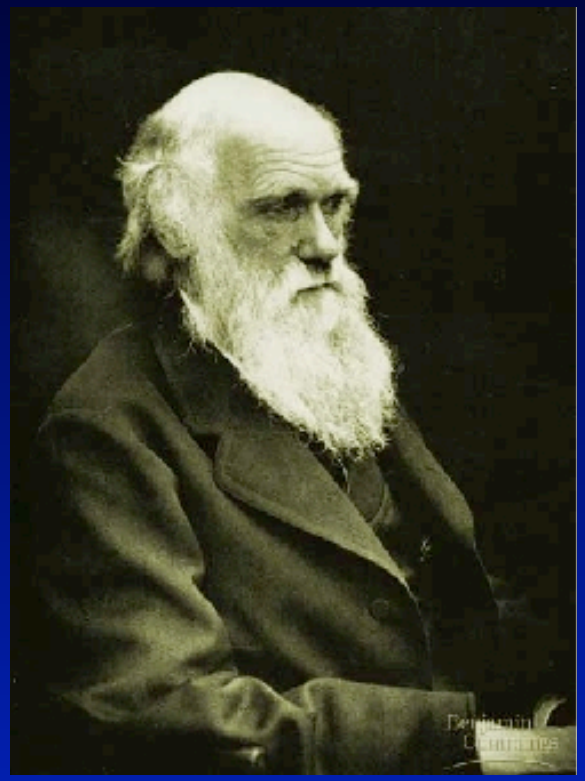
Classification depicted as a tree



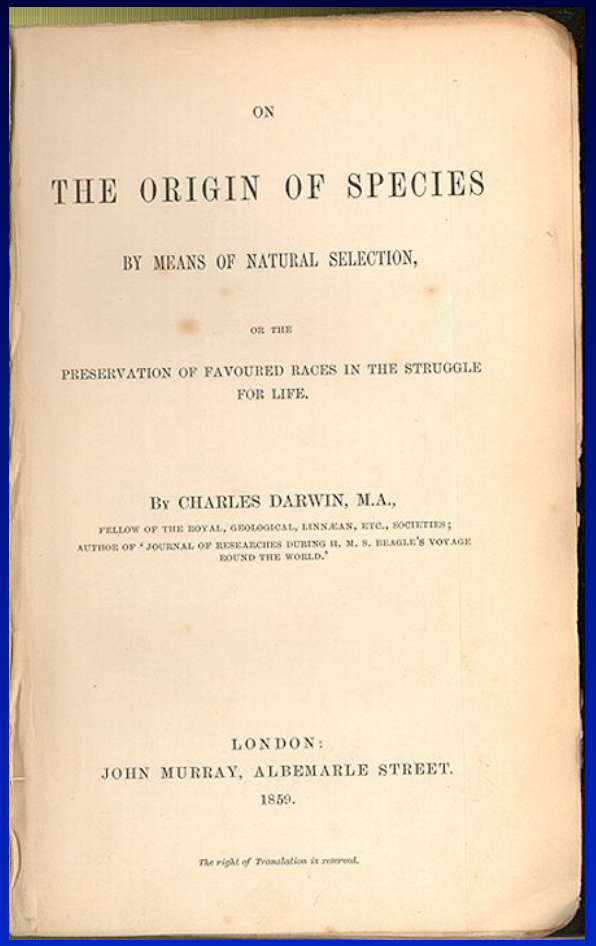
Classification depicted as a tree



Theory of evolution



Charles Darwin
1809-1882



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

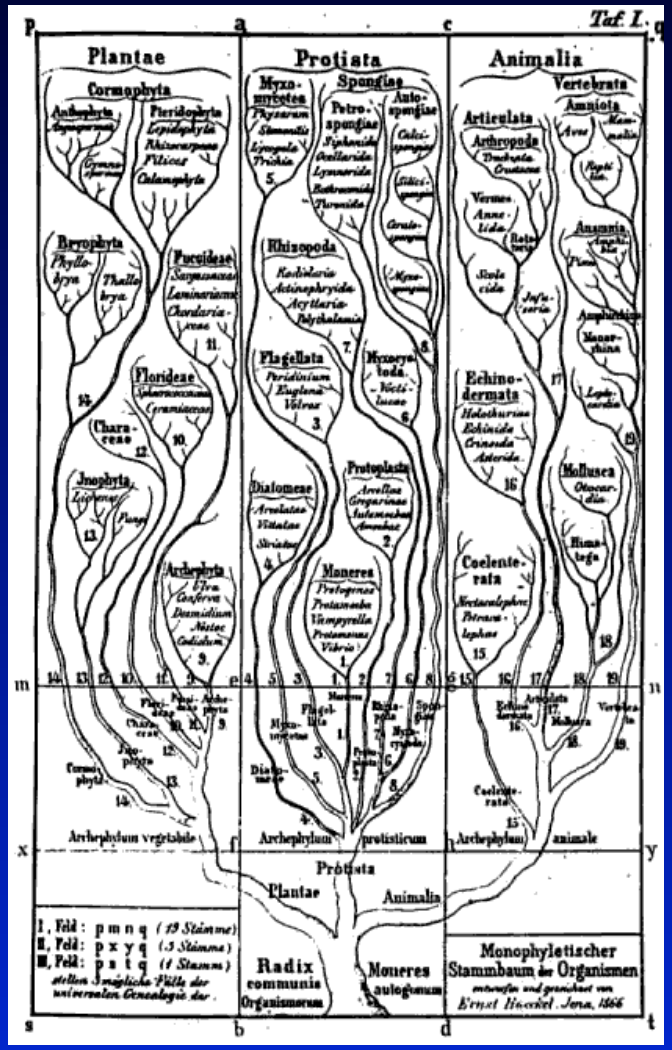
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

Phylogenetic basis of systematics

- Linnaeus:
  Ordering principle is God.
- Darwin:
  Ordering principle is shared descent from common ancestors.
- Today, systematics is explicitly based on phylogeny.

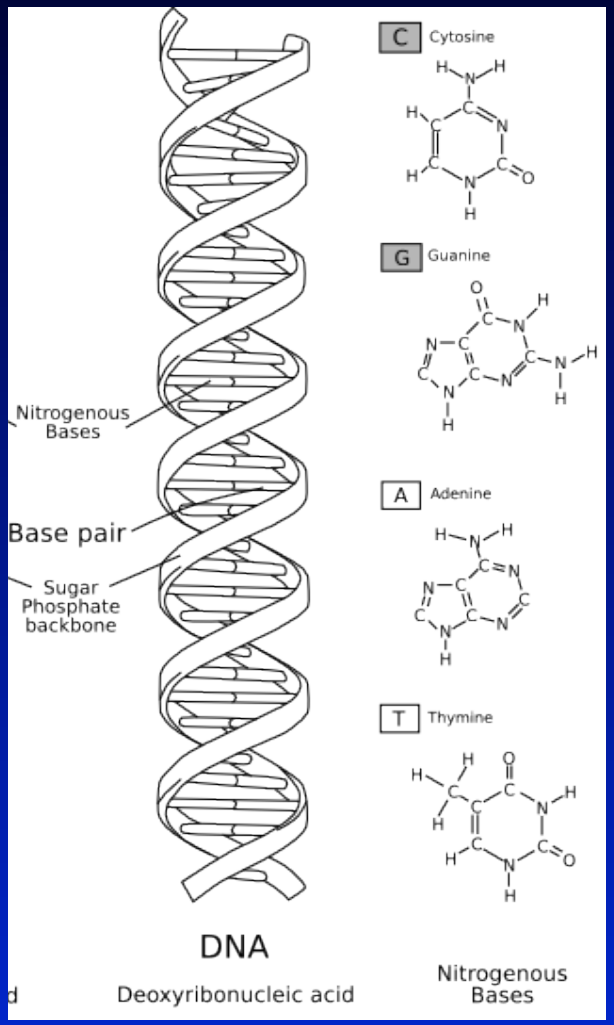


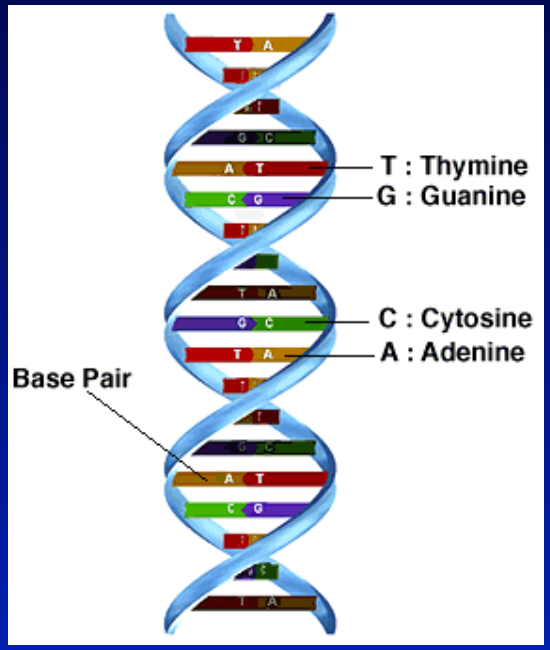
Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.

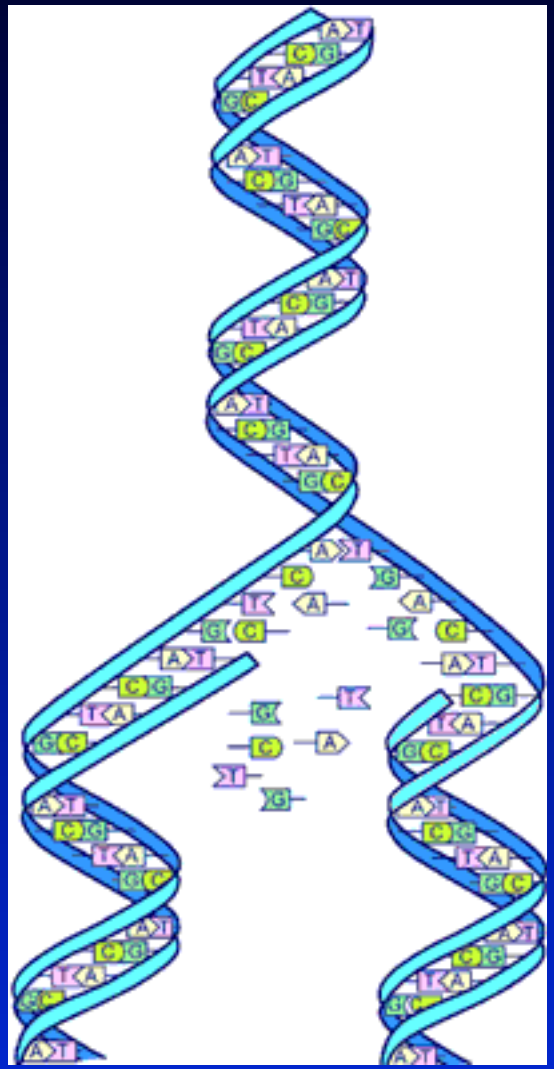
- Evolution by means of natural selection
- Presence of "design-like" features in organisms:
- Quite often features are there "for a reason"

Molecular Basis for Heredity: DNA

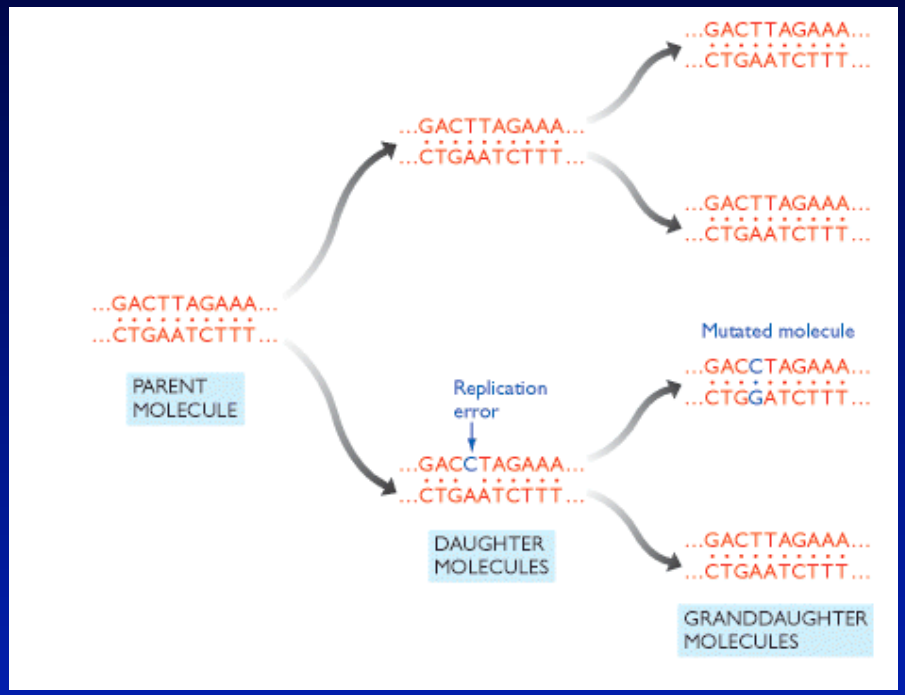




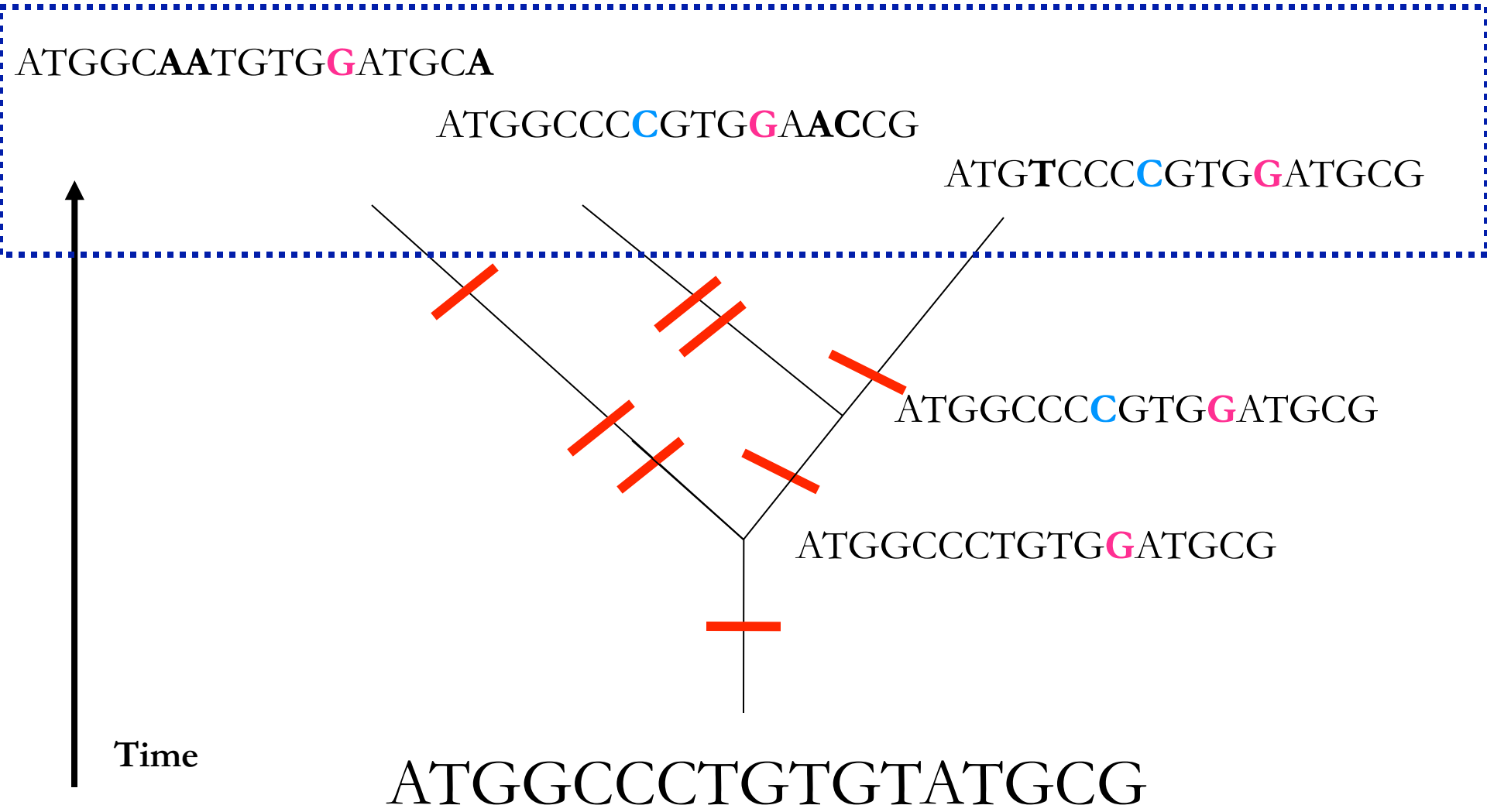
Molecular Basis for Heredity: DNA



Molecular Basis for Variation: DNA Mutation



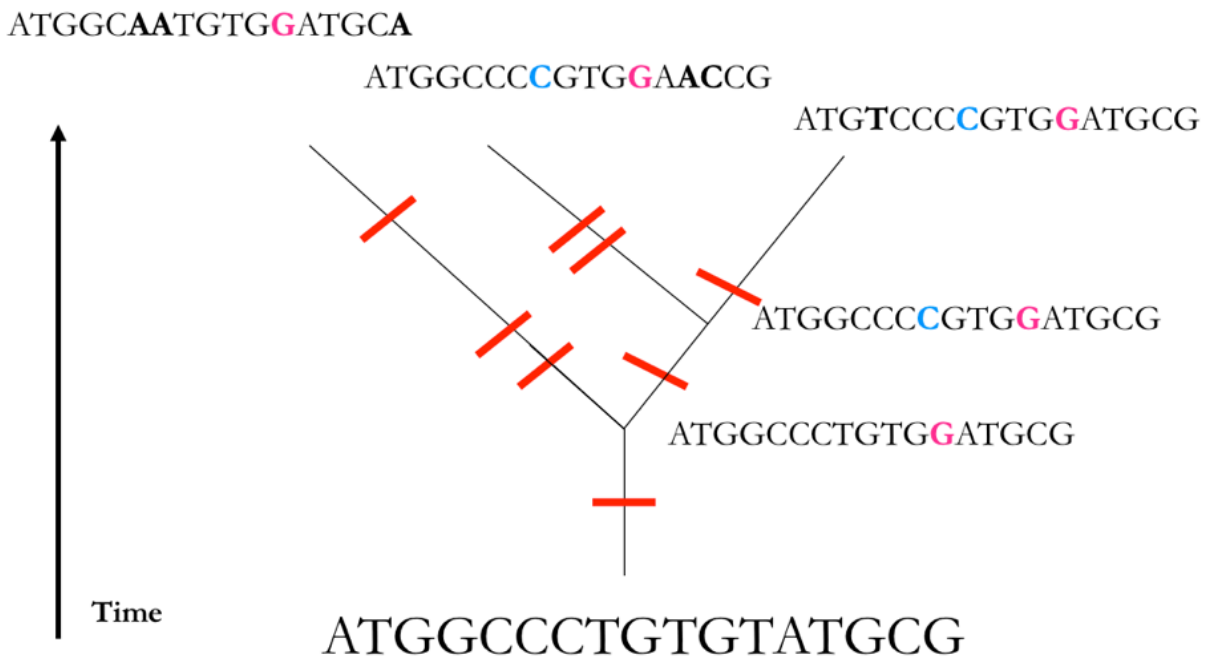
A history of mutations



“DNA alignment”

- Species1: ATGGC**AA**TGTG**G**ATGC**A**
- Species2: ATGGCCC**C**GTG**G**A**AC**CG
- Species3: ATG**T**CCC**C**GTG**G**ATGCG

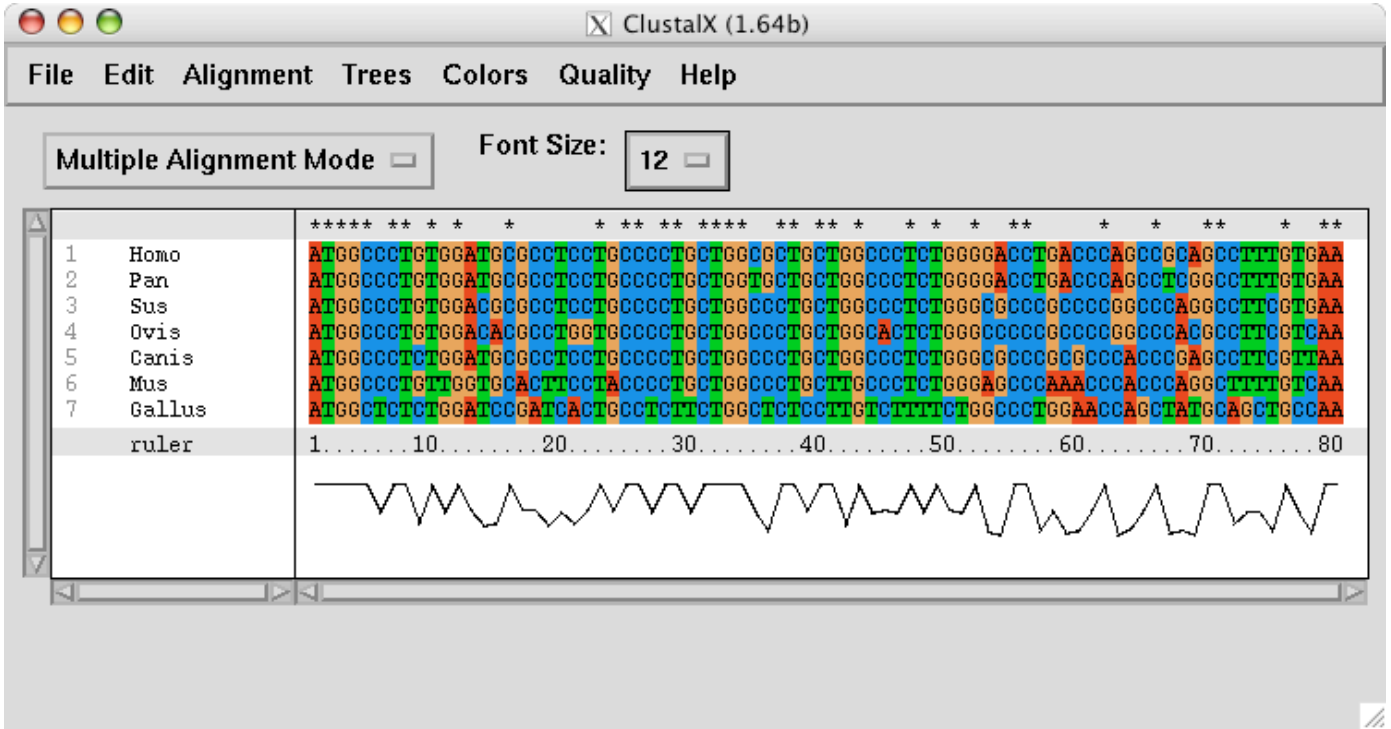
6
3
5



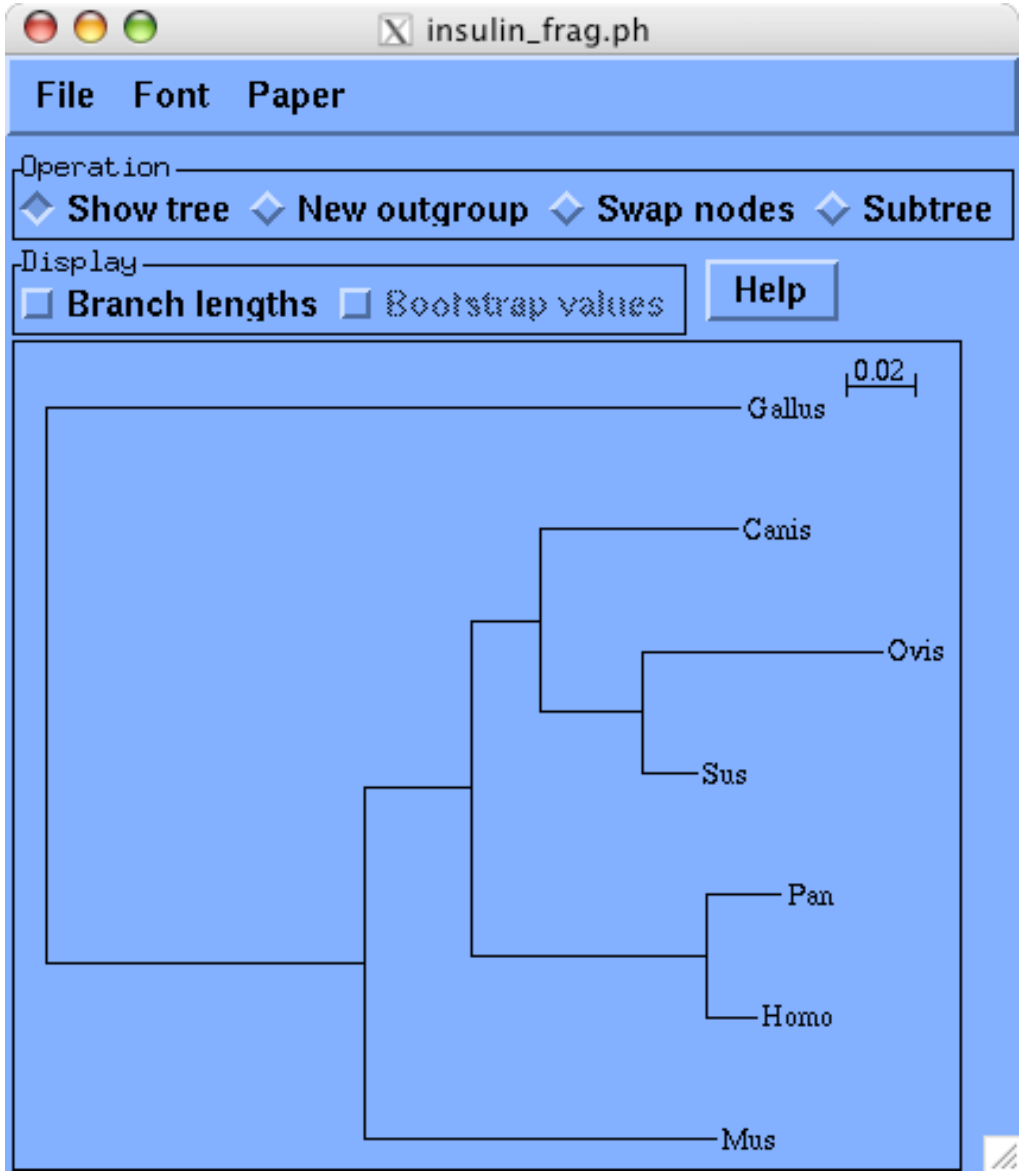
Real life example: Alignment

- Insulin from 7 different species

```
Homo:   ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAA
Pan:    ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGTGCTGCTGGCCCTCTGGGGACCTGACCCAGCCTCGGCCTTTGTGAA
Sus:    ATGGCCCTGTGGACGCGCCTCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCGCCCCGGCCCAGGCCTTCGTGAA
Ovis:   ATGGCCCTGTGGACACGCCTGGTGCCCCTGCTGGCCCTGCTGGCACTCTGGGCCCCCGCCCCGGCCCACGCCTTCGTCAA
Canis:  ATGGCCCTCTGGATGCGCCTCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCGCGCCCACCCGAGCCTTCGTTAA
Mus:    ATGGCCCTGTTGGTGCACTTCCTACCCCTGCTGGCCCTGCTTGCCCTCTGGGAGCCCAAACCCACCCAGGCTTTTGTCAA
Gallus: ATGGCTCTCTGGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGAACCAGCTATGCAGCTGCCAA
```

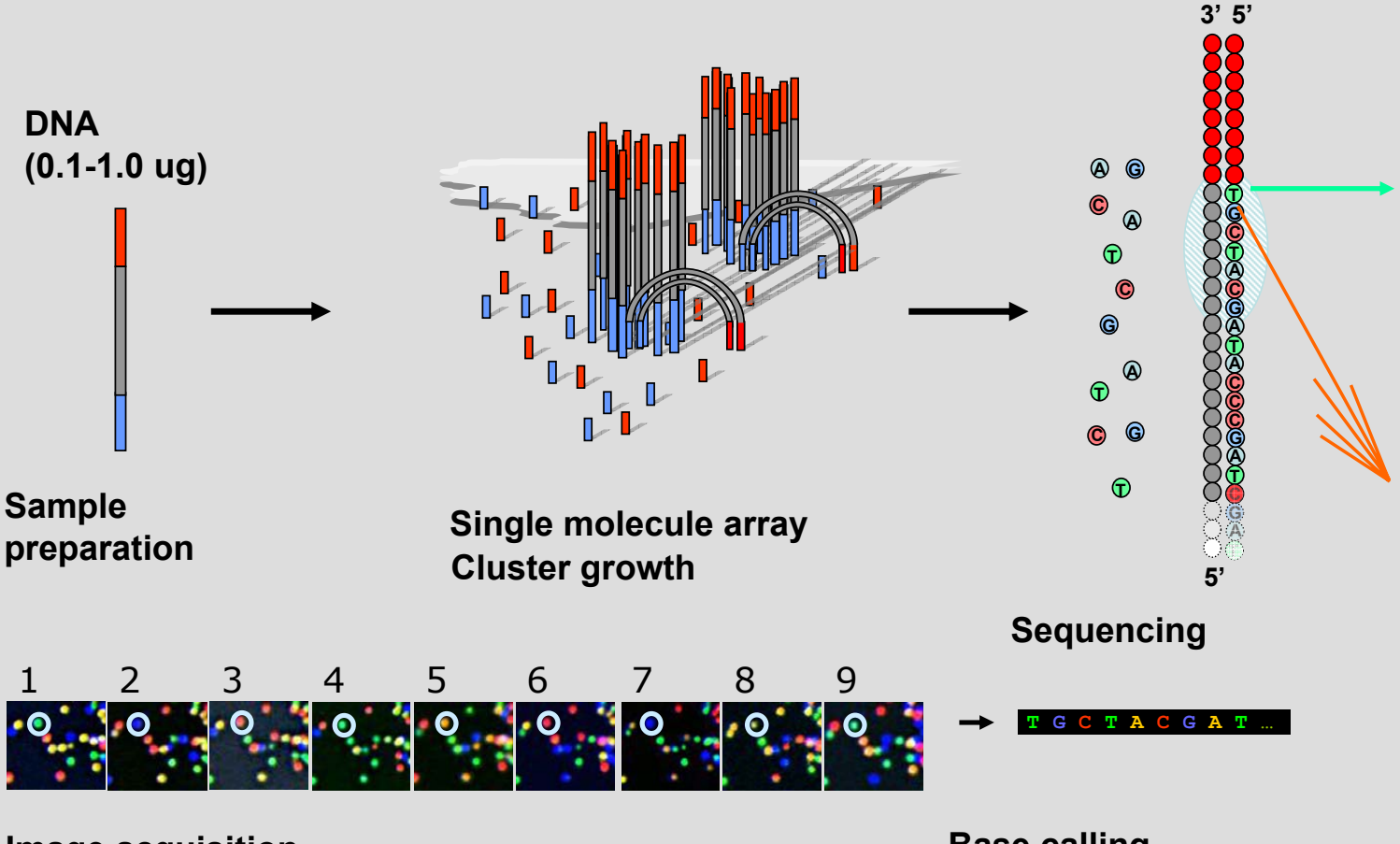


Real life example: Tree

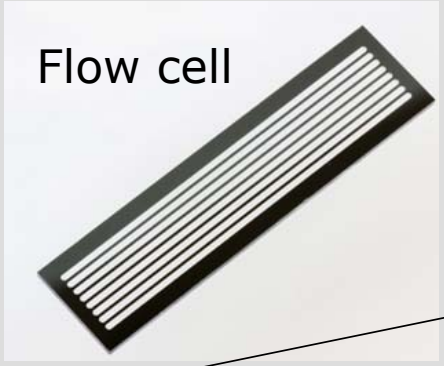


Illumina Sequencing Technology

Reversible Terminator Chemistry

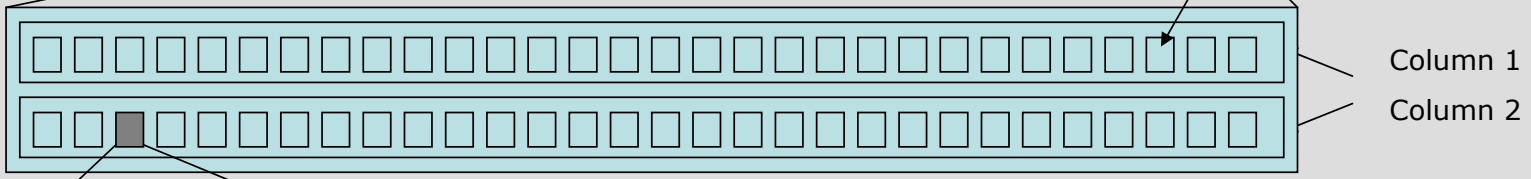


Illumina Sequencing Technology

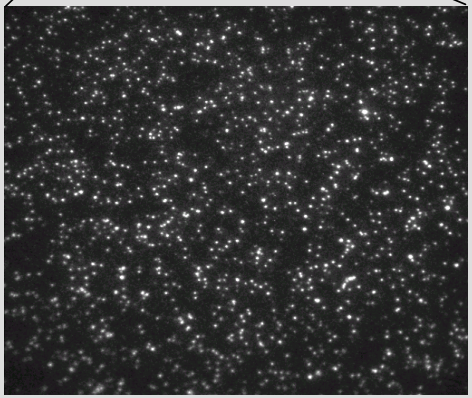


A **flow cell** contains eight lanes

Each **lane** contains **two columns** of tiles



Each **column** contains **multiple tiles** – total 120



Each tile is imaged four times per cycle – one image per base.

~340.000 clusters/tile ->

~40.000.000 clusters/lane ->

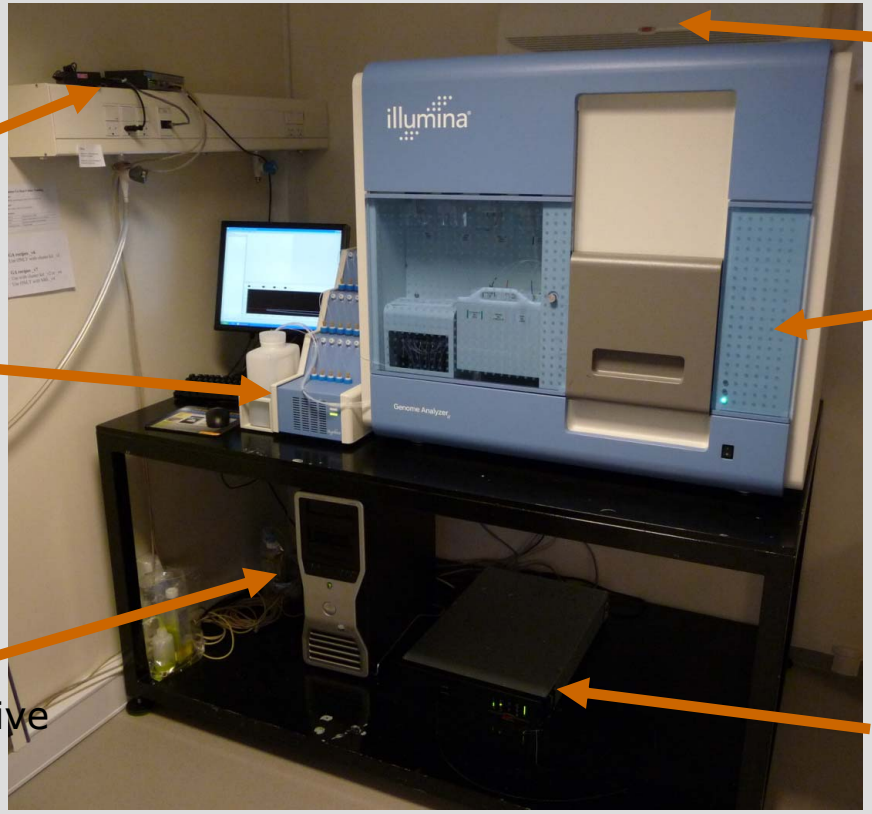
~320.000.000 clusters/flowcell

Illumina GA Sequencing technology

Switch and 100 Mbps network to pipeline computer

Paired End (PE) module

GA PC
- 2.66 GHz cpu
- 3 GB RAM
- 80 GB hard drive



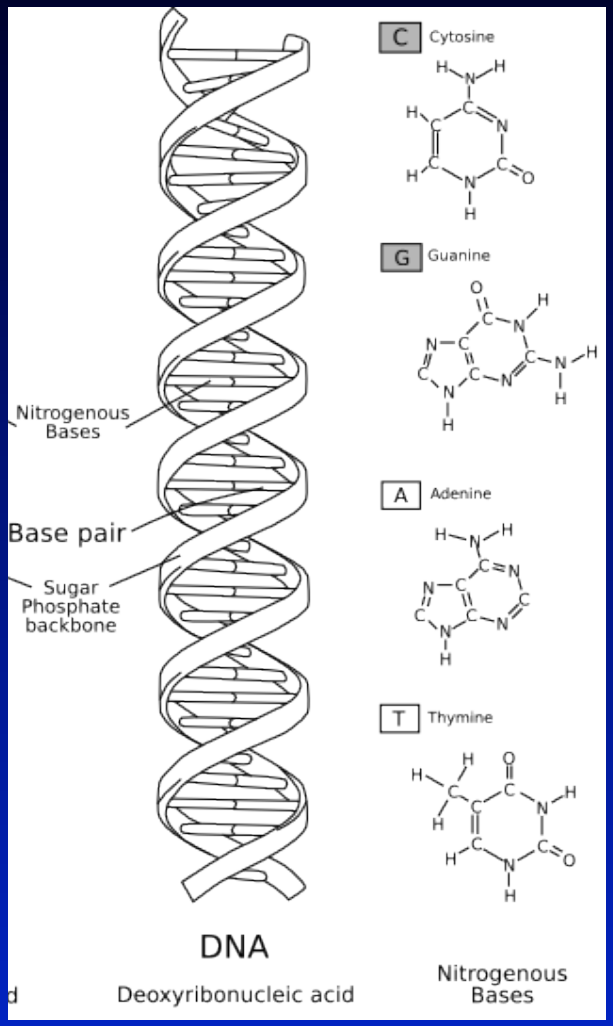
Cooling unit!

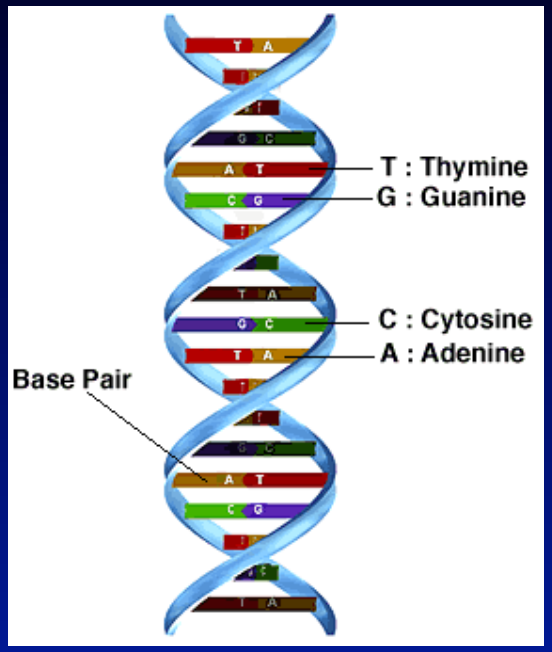
Genome Analyzer (GAIIx)

Uninterruptible Power Supply (UPS)
- Back up for ~10 min

Illumina Genome Analyzer IIx, "Oban"

Symbolic representation of DNA structure

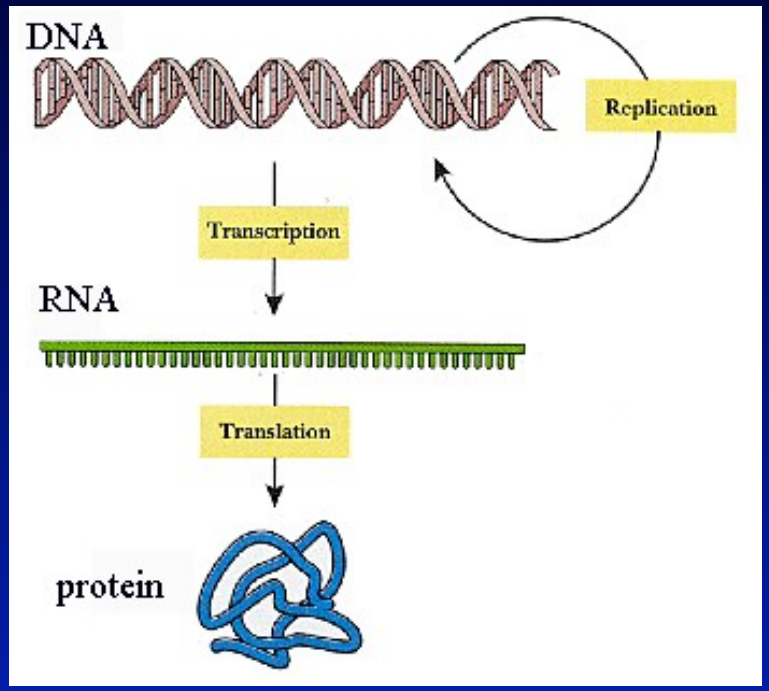




- DNA molecule is a linear polymer
- Structure can be represented as string of 4 symbols: ACTG
- These “sequences” can be analyzed mathematically/ linguistically

HIV genome (approximately 10.000 nucleotides)

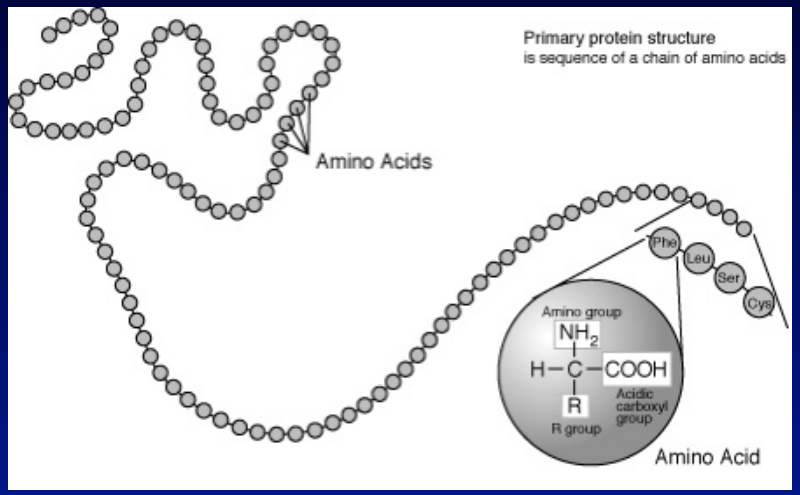
DNA --> RNA --> protein

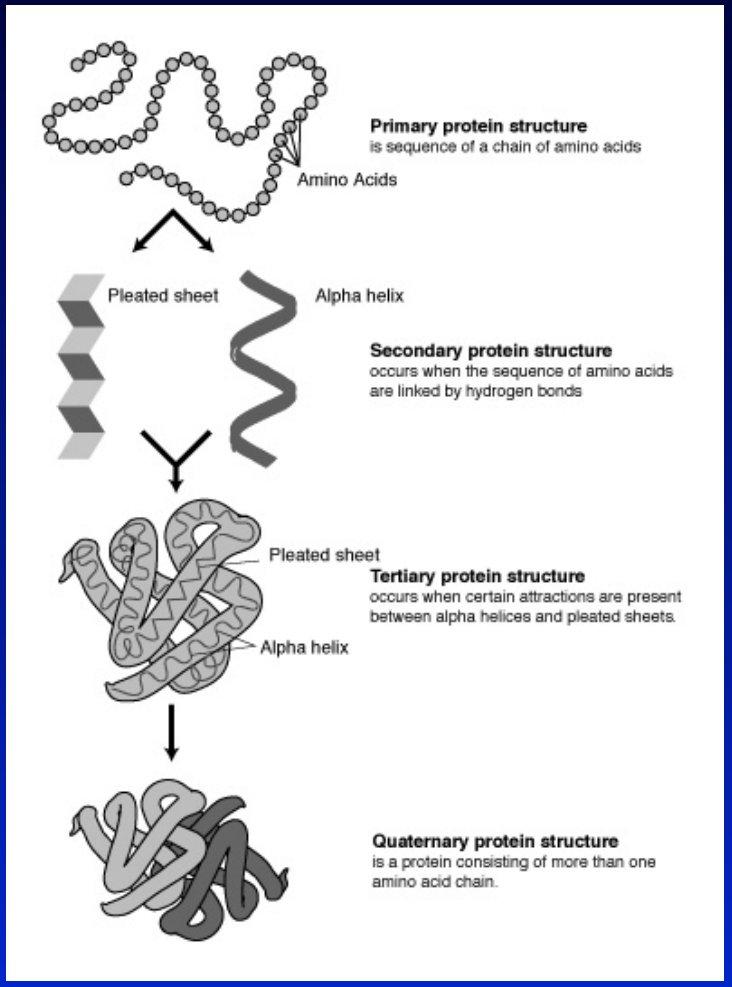


Standard Genetic Code

	T			C			A			G			
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Och	*	TGA	Opa	*	A
	TTG	*Leu*	*L*	TCG	Ser	S	TAG	Amb	*	TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	*Leu*	*L*	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	ATG	*Met*	*M*	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

Symbolic representation of protein structure

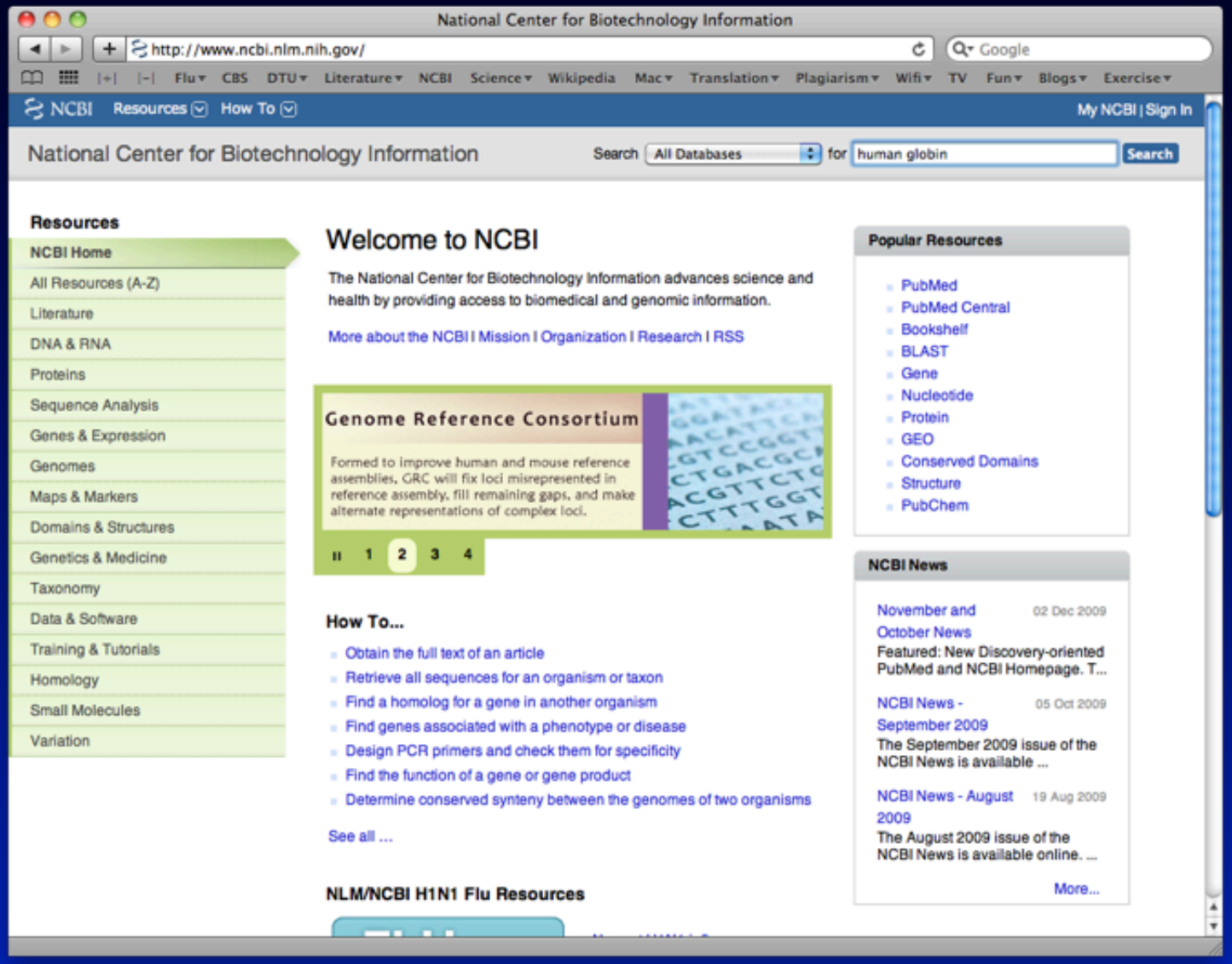




- Proteins are linear polymers
- Built from 20 amino acids
- Can be represented as string of 20 symbols

  ACDEFGHIKLMNPQRSTVWY

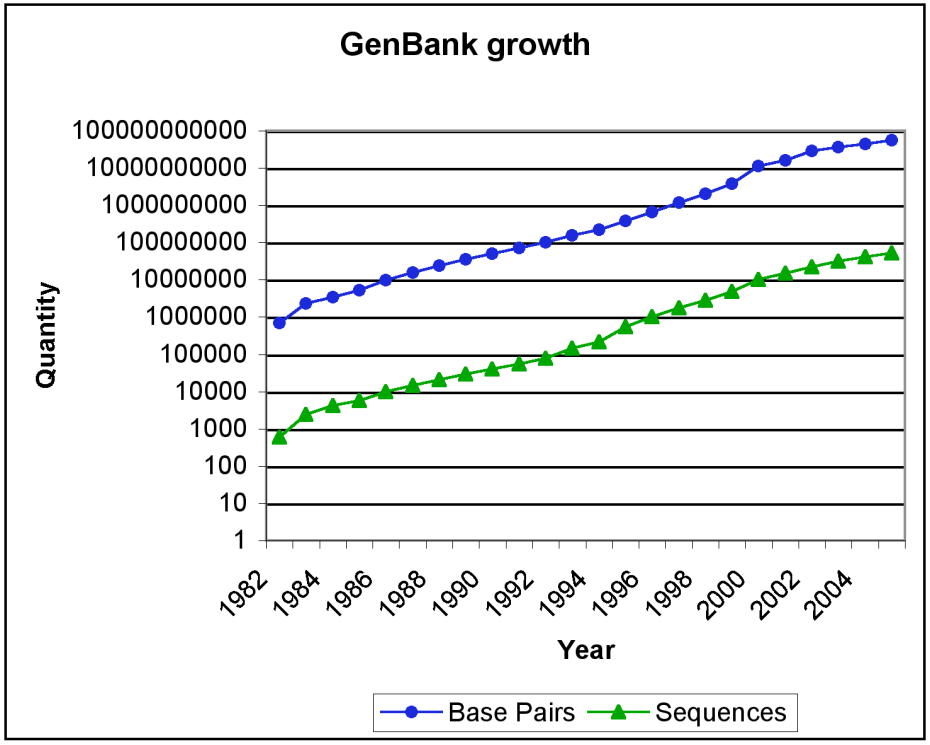
NCBI databases



CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

NCBI GenBank

- GenBank is one of the main international DNA databases.
- GenBank is hosted by NCBI: ***N**ational **C**enter for **B**iotechnology **I**nformation*.
- GenBank has existed since 1982.
- The database is public - no restrictions on the use of the data within.



GenBank format

(handout)

- Originates from the GenBank database.
- Contains both a DNA sequence and annotation of feature (e.g. Location of genes).

GenBank format - HEADER

```
LOCUS       CMGLOAD                 1185 bp    DNA     linear   VRT 18-APR-2005
DEFINITION  Cairina moschata (duck) gene for alpha-D globin.
ACCESSION   X01831
VERSION     X01831.1  GI:62724
KEYWORDS    alpha-globin; globin.
SOURCE      Cairina moschata (Muscovy duck)
  ORGANISM  Cairina moschata
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE   1  (bases 1 to 1185)
  AUTHORS   Erbil,C. and Niessing,J.
  TITLE     The primary structure of the duck alpha D-globin gene: an unusual
            5' splice junction sequence
  JOURNAL   EMBO J. 2 (8), 1339-1343 (1983)
   PUBMED   10872328
COMMENT     Data kindly reviewed (13-NOV-1985) by J. Niessing.
```

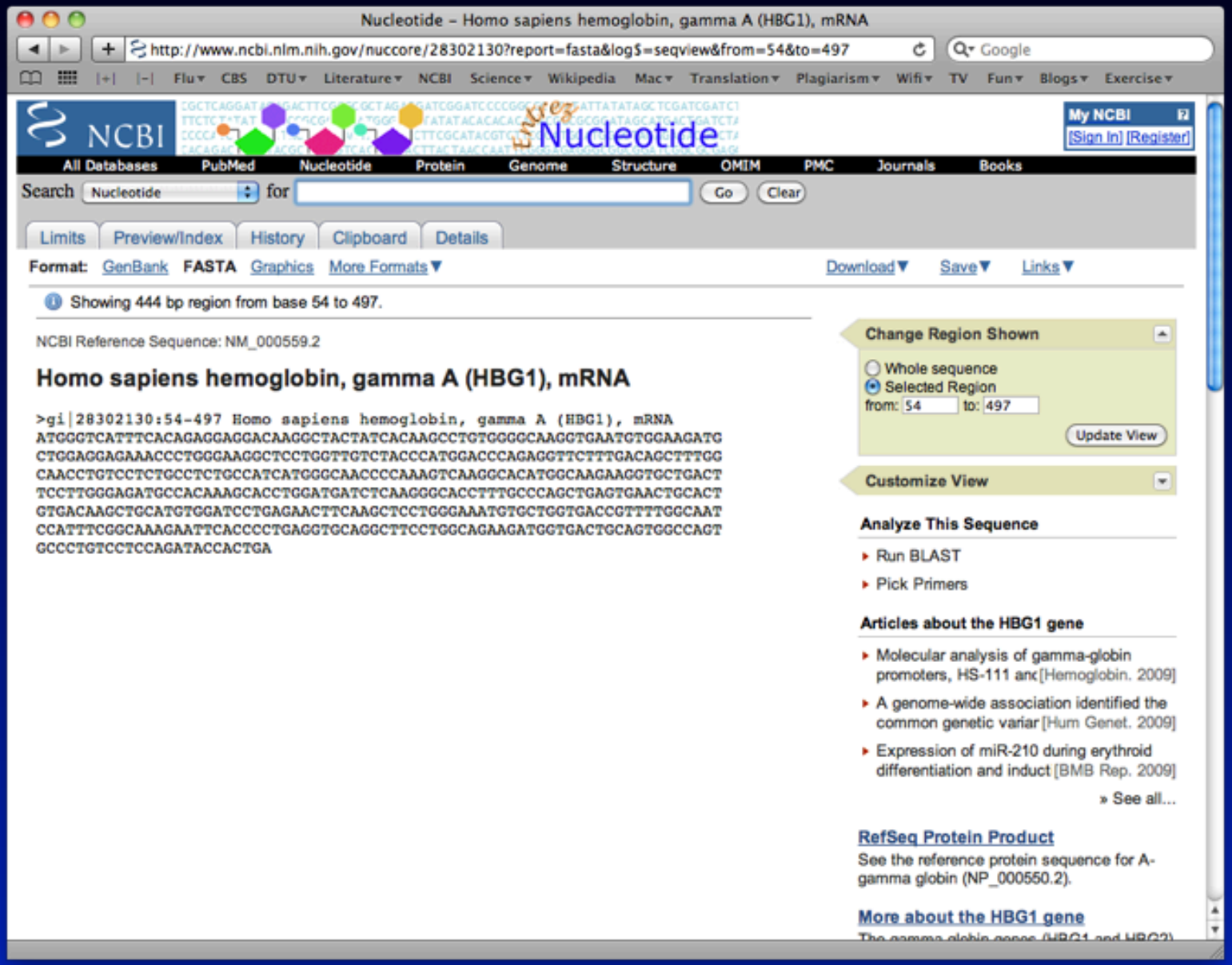
GenBank format - ORIGIN section

```
ORIGIN
        1 ctgcgtggcc tcagcccctc cacccctcca cgctgataag ataaggccag ggcgggagcg
       61 cagggtgcta taagagctcg gccccgcggg tgtctccacc acagaaaccc gtcagttgcc
      121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
      181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
      241 gctgggccca gggggcactc acagggtggg cagcagggag caggagccct gcagcgggtg
      301 tgggctggga cccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
      361 aaaactgact ggcctcgctc cggcaggatg ttcctcgcct acccccagac caagacctac
      421 ttcccccact tcgacctgca tcccggctct gaacaggtcc gtggccatgg caagaaagtg
      481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
      541 ctcagcaacc tgcatgccta caacctgcgt gttgaccctg tcaacttcaa ggcaagcggg
      601 gactagggtc cttgggtctg ggggtctgag ggtgtggggt gcagggtctg ggggtccagg
      661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
      721 gggtaccagg gtcctggggg ccagcagcca gacagcaggg gctgggattg catctgggat
      781 gtgggccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
      841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
      901 tccggagcag gggtactaag ccctggtttg ccttgcagct gctggcacag tgcttccagg
      961 tggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
     1021 agttcttgtc cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
     1081 cccttgcacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
     1141 gggcatcggg ggtcccaggg agggctgggt tgcttccaca catcc
//
```

GenBank format - FEATURE section

```
FEATURES             Location/Qualifiers
     source          1..1185
                     /organism="Cairina moschata"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:8855"
     CAAT_signal     20..24
     TATA_signal     69..73
     precursor_RNA   101..1114
                     /note="primary transcript"
     exon            101..234
                     /number=1
     CDS             join(143..234,387..591,939..1067)
                     /codon_start=1
                     /product="alpha D-globin"
                     /protein_id="CAA25966.2"
                     /db_xref="GI:4455876"
                     /db_xref="GOA:P02003"
                     /db_xref="InterPro:IPR000971"
                     /db_xref="InterPro:IPR002338"
                     /db_xref="InterPro:IPR002340"
                     /db_xref="InterPro:IPR009050"
                     /db_xref="UniProt/Swiss-Prot:P02003"
                     /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
                     HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLA
                     QCFQVVLAAHLGKDYSPEMHAAFDKFLSAVAAVLAEKYR"
     repeat_region   227..246
                     /note="direct repeat 1"
     intron          235..386
                     /number=1
     repeat_region   289..309
                     /note="direct repeat 1"
     exon            387..591
                     /number=2
     intron          592..939
                     /number=2
     exon            940..1114
                     /number=3
     polyA_signal    1095..1100
     polyA_signal    1114
```

NCBI databases: fasta format



FASTA file